А.Г. Буховец
(*Воронеж*)

ОБ ОДНОМ ПОДХОДЕ К ЗАДАЧЕ КЛАССИФИКАЦИИ

Предлагаемый подход к задаче многомерной классификации (кластерного анализа) позволяет при довольно общих предположениях объяснить механизм формирования структуры данных. Приводятся модельные примеры, демонстрирующие возможности этого подхода. Рассматривается содержательный пример классификационной задачи, связанный с распределением среднедушевого дохода населения России.

Ключевые слова: методы многомерной классификации, кластерный анализ, типологический анализ, моделирование структуры многомерных данных, уравнение Шредингера.

Проблема классификационной задачи (типологического анализа) в социологических исследованиях имеет давно сложившиеся традиции в постановках и подходах к решению. Результаты, достигнутые в этом направлении, были обобщены и систематизированы в ряде монографий и статей, посвященных этой тематике (см., например, [1; 2; 3; 4; 5]).

К концу 80-х гг. стало понятно, что дескриптивный (описательный) подход к построению классификационных моделей себя в основном уже исчерпал. Об этом можно судить хотя бы потому, что практически все современные профессиональные системы статистической обработки данных, например, такие как STATICTICA, STATGRAPHICS, SPSS, включают в себя примерно

Алексей Георгиевич Буховец – кандидат экономических наук, доцент Воронежского государственного агроуниверситета.

один и тот же набор алгоритмов кластерного анализа, созданных в предыдущие десятилетия.

Анализируя опыт построения классификаций в социологических исследованиях [1; 4] и развивая идею исследования структуры многомерных данных, можно прийти к пониманию того, что следующий шаг в дальнейшем развитии этого направления должен заключаться в изучении механизмов формирования структуры многомерных данных.

Предлагаемый нами подход следует именно этому направлению. В его основу положены два следующих наиболее существенных принципа. Первый из них касается отказа от изотропности признакового пространства. Это достигается путем введения некоторого потенциала, изменяющего топологию пространства. В рассматриваемых ранее классификационных моделях практически никогда не вводилось каких-либо ограничений на значения признаков, хотя каждому специалисту, занимающемуся обработкой социологических данных, хорошо известно, что эти значения всегда имеют практические границы.

Второй аспект данного подхода связан с предположением о характере движения объектов в признаковом пространстве. Предполагается, что изменение положения в признаковом пространстве не вызывает нарушения структуры данных и, как показано, может быть описано соответствующим дифференциальным уравнением. При этом ограничения, записанные в дифференциальной форме, могут быть проинтегрированы и связи из кинематических переведены в статистические. Эти связи как раз и будут соответствовать структуре многомерных данных.

Логическое обоснование формальной постановки классификационной задачи

Задача классификации объектов произвольной природы на понятийном уровне может быть сформулирована как задача разделения совокупности объектов на некоторые классы (группы, кластеры), являющиеся однородными в некотором смысле. Такая постановка задачи, сводящаяся к выделению классов однородных объектов, встречается в той или иной степени практически в любой отрасли человеческой деятельности. Сложность, характеризующая степень неопознанности объекта, приводит к тому, что часто очень трудно построить классификацию всей совокупности объектов, используя только информацию одного какого-либо признака.

В условиях типологического анализа исследователь обычно располагает некоторым набором признаков, которые, как он считает, позволяют описывать сущностные свойства объектов. В этом случае изучение реальных объектов заменяется изучением образов объектов, представленных набором признаков, характеризующих реальный объект. При этом имплицитно имеет место предположение, что объекты, схожие с точки зрения изучаемого процесса, будут обладать набором примерно одинаковых по значению признаков. И напротив, различие в значениях признаков может быть проявлением того, что объекты являются качественно разнородными, т.е. принадлежат к разным классам. Таким образом, качественная однородность объектов и, следовательно, принадлежность их к одному и тому же классу ставится в зависимость от значений наблюдаемых признаков. В связи с этим возникает проблема численной оценки степени сходства объектов, т.е. задача перехода от качественного уровня измерения сходства к измерению степени сходства в шкалах более высокого порядка.

Наиболее просто эта проблема находит свое разрешение в рамках геометрического подхода, суть которого заключается в следующем. Совокупность признаков, характеризующих рассматриваемые объекты, образует признаковое пространство, в котором каждый объект представлен в виде точки. Тогда наличие областей признакового пространства с более высокой концентрацией точек позволяет говорить о существовании некоторых классов объектов. Задача классификации при таком подходе может быть сформулирована как задача исследования структуры многомерных данных в признаковом пространстве [1].

Заметим, что процедура классификации, имеющая цель получить содержательно интерпретируемое разбиение множества объектов, т.е. типологию, почти никогда не применяется к произвольной совокупности объектов. Как правило, рассматриваемая совокупность обладает системными свойствами и исследуется как некоторая система. В этом случае взаимодействие объектов с внешней средой и между собой объясняет возникновение системных свойств и наличие структуры системы. Учитывая, что элементы системы – образы реальных объектов – представлены в виде точек пространства и рассматриваются как математические объекты, не обладающие собственной внутренней структурой, будет естественным предположить, что для описания их взаимодействия необходимо ввести некоторый потенциал $U = U(x)$, характеризующий неоднородность признакового пространства. Посредством этого потенциала система оказывает некоторое воздействие на объекты, т.е. реагирует на происходящие изменения. Все это позволяет сделать предположение о том, что в признаковом пространстве определен некоторый потенциал $U = U(x)$, являющийся функцией координат точек пространства.

Как известно, в традиционной постановке задачи кластерного анализа рассматриваются признаки объектов, которые не зависят от времени. (Мы не имеем в виду случаи, когда производится классификация траекторий объектов в признаковом пространстве.) Однако предположение о том, что положение точек в пространстве допускает некоторые локальные изменения, не противоречит сути классификационной задачи. Возможность существования локальных изменений может быть допущена хотя бы в силу наличия ошибок измерения признаков. Кроме этого, локальные изменения, очевидно, могут быть вызваны эволюцией самого объекта. Но в любом случае эти изменения не должны затрагивать глобальной структуры многомерных данных.

Допустив существование локальных изменений, мы неявным образом вводим некоторые динамические характеристики объектов в признаковом пространстве.

Очевидно, и это следует из практики, что эти локальные изменения не должны быть столь большими, чтобы быть в состоянии нарушить фиксируемую (оцениваемую) структуру данных. Следовательно, их величина должна быть ограничена некоторыми внешними по отношению к ним силами, удерживающими значения признаков от произвольно больших изменений. Отказ от этого предположения означал бы, что мы не можем наблюдать никаких фиксированных состояний классов объектов, а вынуждены лишь констатировать произвольные, хаотические изменения значений признаков. Введенный ранее в рассмотрение потенциал $U(x)$ позволяет избежать такой ситуации.

Таким образом, задание потенциала в признаковом пространстве может рассматриваться в качестве основной причины формирования системы объектов. Очевидно, что именно этот потенциал и будет отвечать за формирование структуры данных в признаковом пространстве.

Природа рассматриваемого потенциала может быть, на наш взгляд, самой различной. В качестве определения потенциала, например, может выступать описание воздействия внешней по отношению к системе объектов среды. В этом случае структура системы будет представлять собой реакцию функционирующей системы на воздействия внешней среды. Очевидно, что изменения потенциала будут приводить к соответствующим изменениям в структуре системы.

В качестве примера проявления свойств потенциала сравним классовую структуру современного общества в России и, например, СССР периода до 1985 г. Со всей очевидностью можно отметить существование значительных различий. О причинах, приведших к изменению структуры классов, можно сказать почти однозначно: это изменение политической ориентации страны, повлекшее за собой радикальные изменения в экономике и связанными с ней социальными преобразованиями. Таким образом, изменение социально-экономических условий в стране привело к существенному изменению классовой структуры общества. Снятие ограничений на виды деятельности способствовало как появлению новых классов, так и изменению состава ранее существовавших. Если считать, что виды деятельности и способы жизнеобеспечения за этот период значительно не изменились, то функцию потенциала следует определить как некоторую интегральную характеристику, отражающую социально-экономические условия существования общества.

Вторым важным аспектом, отвечающим за формирование структуры данных, будет закон движения объектов в признаковом пространстве с заданным потенциалом. Поскольку этот закон отражает динамические характеристики объектов, то желательно представить его в виде некоторого дифференциального уравнения. Решение такого дифференциального уравнения должно приводить к описанию наблюдаемой кластерной структуры многомерных данных.

Классификация как задача о собственных значениях

Предположим, что имеется система объектов, представленных в виде точек n-мерного признакового пространства. Кроме этого будем предполагать, что в признаковом пространстве определен некоторый потенциал $U = U(x)$, где через x обозначены координаты точки n-мерного признакового пространства. Распределение точек в признаковом пространстве будет характеризоваться некоторой, вообще говоря, комплексно-значной функцией $\psi(x,t)$. При этом величина $\psi^2(x,t)$ будет численно равняться плотности вероятности распределения объектов в точке x в момент времени t.

Пусть x_0 – точка пространства, соответствующая некоторому объекту системы. Допустим, что при малых отклонениях объекта от положения x_0 в системе возникает некоторая сила, препятствующая этому отклонению. Тогда, если в признаковом пространстве выделить некоторый элементарный объем в виде куба с ребром Δx_i и центром в точке x_0, то, как известно, разность между средним значением функции $\overline{\Psi}$, вычисленная по рассматриваемому объему куба, и значением $\Psi_0 = \Psi(x_0)$ может быть представлена в следующем виде:

$$\overline{\Psi} - \Psi_0 = C\sum_i^n \frac{\partial^2 \Psi}{\partial x_i^2} \equiv C\nabla^2\Psi,$$

где C – некоторая константа, определяемая размерностью пространства n и величиной Δx_i.

Предположим, что величина этого отклонения компенсируется действием некоторой внутренней силы, удерживающей систему в стабильном состоянии, т.е. реакцией системы на изменения значений признаков объекта. Пусть эта величина может быть представлена в виде $(E - U)\psi$, где $U = U(x)$ – ранее введенный потенциал, отвечающий за сохранение системы, E – величина, играющая роль энергии движущейся точки. Введенное в рассмотрение соотношение означает, что энергия движущейся точки уменьшается действием сдерживающего потенциала $U(x)$.

Итак, изменение положения точки в признаковом пространстве может быть представлено в виде суммы двух слагаемых

$$C\nabla^2\Psi + (E - U)\Psi = 0$$

и выражает условие сохранения системы в пространстве.

Полученное соотношение характеризует изменение величины функции ψ, которое вызвано смещением объекта, находящегося в точке x_0, за счет изменения координат точки в признаковом пространстве. Однако при этом следует принять во внимание, что при изменении положения объекта в признаковом пространстве может меняться не только абсолютная величина смещения, но и направление смещения. Будем считать, что поворот вектора, характеризующего направление смещения, может быть учтен введением сомножителя $e^{-i(E-U)t}$, который, очевидно, не изменяет абсолютной величины смещения. В этом случае функция $\psi(x,t) = \varphi(x)e^{-i(E-U)t}$ будет представлена в виде произведения двух сомножителей, один из которых связан с положением точки в пространстве и зависит только от x, а другой характеризует изменение смещения, вызванное реакцией системы на происходящее изменение. Тогда изменение функции ψ, вызванное изменением положения объекта во времени, можно представить так:

$$\frac{\partial\Psi}{\partial t} = \frac{\partial}{\partial t}(\varphi(x)e^{-i(E-U)t}) = -i(E-U)\varphi(x)e^{-i(E-U)t} = -i(E-U)\Psi$$

или

$$\frac{\partial\Psi}{\partial t} + i(E-U)\Psi = 0.$$

Умножая обе части соотношения на i, получим:

$$i\frac{\partial\Psi}{\partial t} - (E-U)\Psi = 0.$$

Полученное равенство выражает условие сохранения системы во времени.

Полагая, что изменение состояния системы в признаковом пространстве соответствует изменению системы во времени, получим уравнение, описывающее состояние системы в точке x

$$i\frac{\partial\Psi}{\partial t} = C\nabla^2\Psi + 2(E-U)\Psi.$$

Таким образом, локальное состояние системы в признаковом пространстве описывается операторным уравнением:

$$i\frac{\partial\Psi}{\partial t} = H\Psi,$$

где $H = C\nabla^2 + 2(E-U)$ – оператор.

Уравнение такого вида известно как уравнение Шредингера [6]. Для его решения может быть использован метод разделения переменных. Полагая

$$\Psi(x,t) = \varphi(x)f(t),$$

получим

$$i\frac{\partial f(t)}{\partial t}\varphi(x) = H\varphi(x)f(t).$$

Разделим переменные x и t, и, обозначая константу разделения w, будем иметь

$$i\frac{\partial f(t)}{\partial t} = wf(t).$$

С учетом того, что функция f зависит только от t, имеем:

$$f(t) = Ae^{-iwt},$$

где A – константа интегрирования.

Подставляя найденное выражение для $f(t)$ в операторное уравнение, получим

$$H\varphi(x) = w\varphi(x).$$

Таким образом, решение уравнения свелось к задаче определения собственных функций $\psi_n(x)$ и собственных значений w_n оператора H. Введенный в рассмотрение оператор H является самосопряженным, а следовательно, имеет дискретный спектр действительных собственных значений w_n, а его собственные функции $\psi_n(x)$ – ортогональны. Поэтому общее решение уравнения может быть представлено в виде ряда по собственным функциям $\psi_n(x)$. После соответствующей нормировки величина $|\psi|^2$ будет численно равна плотности вероятности распределения объектов системы в состоянии, характеризуемом набором признаков x. Это позволяет сравнивать результаты работы алгоритма кластерного анализа со сделанными выше предположениями.

Модельные примеры, демонстрирующие возможности предлагаемого подхода

Для того чтобы продемонстрировать возможности предлагаемого подхода к решению классификационной задачи, рассмотрим в качестве модели функции распределения объектов в одномерном случае в поле потенциала, не зависящего от времени. Такая постановка задачи приводит к одномерному стационарному уравнению Шредингера. В этом случае, как было отмечено ранее, функция может быть представлена в виде произведения двух функций, одна из которых зависит только от координат, а другая представляет собой осциллирующую функцию, зависящую от величины энергии E.

Решения полученного уравнения для потенциалов вида $U(x) = Cr^n$ приводятся в литературе. Хорошо известны решения и названия потенциалов для значений n, равных –2, –1, 0, 1, 2. В этих случаях уравнение Шредингера сводится к обыкновенному дифференциальному уравнению второго порядка, решение которого может быть выражено через известные элементарные и специальные функции. Качественный анализ этих решений показывает, что в случае задания ограничений на движение объектов хорошо воспроизводится такая структурная особенность данных, как наличие отдельных устойчивых состояний. Однако с точки зрения классификационного подхода, полученные результаты не могут считаться удовлетворительными, поскольку при задании потенциалов такой формы происходит совмещение максимумов функций распределений, соответствующих различным классам. Проблема, на наш взгляд, может быть разрешена путем введения потенциала, не обладающего симметрией, т.е. за счет деформации формы потенциала.

В качестве модели классификационной задачи рассмотрим поведение объектов, описываемое стационарным уравнением Шредингера в бесконечно глубокой потенциальной яме. Этому случаю соответствует дифференциальное уравнение второго порядка

$$-\frac{d^2\Psi}{dX^2} + U\Psi = E\Psi$$

где значение величины C для удобства исследования будем считать равным единице.

Для получения решений перейдем к рассмотрению сеточных функций $\Psi_n = \Psi(x_n)$, разбив интервал [0, L], на котором определен потенциал $U(x)$ на N интервалов точками $x_n = n\bullet\Delta$, где $n = 1, 2, \dots N$,

$\Delta = L/N$. Обозначим $U_n = U(x_n)$. Заменяя дифференциальное уравнение конечно-разностным [7], будем иметь

$$\tfrac{1}{\Delta^2}\left[\Psi_{n-1} - 2\Psi_n + \Psi_{n+1}\right] + U_n\Psi_n = E\Psi_n.$$

Учитывая граничные условия $\Psi_0 = \Psi_N = 0$, получим систему однородных линейных уравнений, которую можно представить в матричном виде

$$\mathbf{G}\Psi = \mathbf{E}\Psi,$$

где

$$\mathbf{G} = \begin{pmatrix} \frac{2}{\Delta^2}+U_1 & -\frac{1}{\Delta^2} & 0 & 0 & 0 & \ldots & 0 \\ -\frac{1}{\Delta^2} & \frac{2}{\Delta^2}+U_2 & -\frac{1}{\Delta^2} & 0 & 0 & \ldots & 0 \\ 0 & -\frac{1}{\Delta^2} & \frac{2}{\Delta^2}+U_3 & -\frac{1}{\Delta^2} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & 0 & 0 & \ldots & -\frac{1}{\Delta^2} & \frac{2}{\Delta^2}+U_{N-1} & -\frac{1}{\Delta^2} \\ 0 & 0 & 0 & \ldots & 0 & -\frac{1}{\Delta^2} & \frac{2}{\Delta^2}+U_N \end{pmatrix}$$

$$\Psi = \left(\psi_1, \quad \psi_2, \quad \ldots, \quad \psi_{N-1}, \quad \psi_N\right)^T$$

Полученная однородная система N линейных уравнений будет иметь нетривиальные решения только при значениях E_n ($n = 1, \ldots, N$), являющихся собственными значениями матрицы $\mathbf{G}$. Таким образом, задача о распределении объектов в одномерном признаковом пространстве с заданным потенциалом $U(x)$ свелась к задаче нахождения собственных векторов $\Psi^{E_n} = \{\Psi_n\}$ матрицы $\mathbf{G}$.

На приведенном ниже рис. 1 представлены результаты численных расчетов, выполненных для потенциала бесконечно глубокой ямы с рельефом дна, задаваемым функцией $U(x)$. На этом рисунке изображены функции распределений для первых четырех классов, т.е. значения квадратов сеточных функций

$(\Psi^{(1)})^2, (\Psi^{(2)})^2, (\Psi^{(3)})^2, (\Psi^{(4)})^2$. Для сравнения здесь же приведена сеточная функция потенциала U_n. Все графики функций $(\Psi^{(n)})^2$ располагаются на уровне потенциала, значение которого равняется соответствующему собственному значению E_n. Отличительной особенностью примера является унимодальность полученных решений, а также довольно хорошее различение максимумов всех функций распределений.

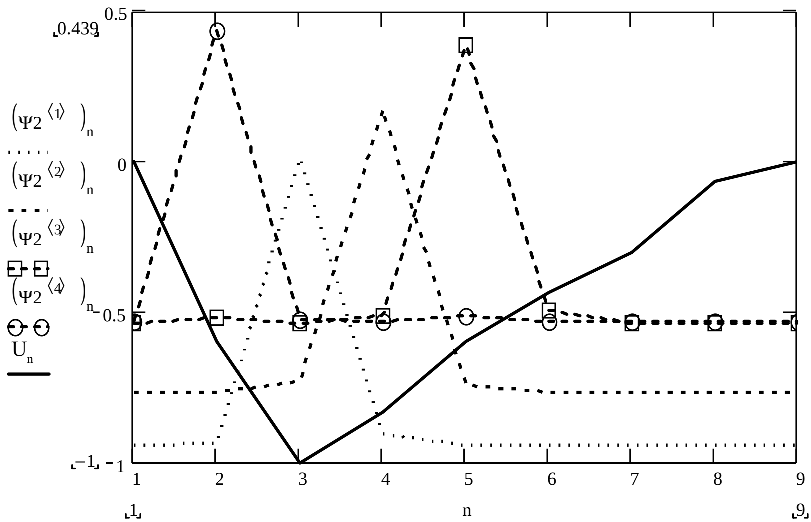

Рис. 1. Квадраты собственных функций для первых четырех собственных значений бесконечно глубокой потенциальной ямы с несимметричным рельефом дна

Аналогичные результаты можно видеть на рис. 2. Сравнение результатов этих расчетов позволяет составить некоторое представление о влиянии формы потенциала на расположение мод функций распределений.

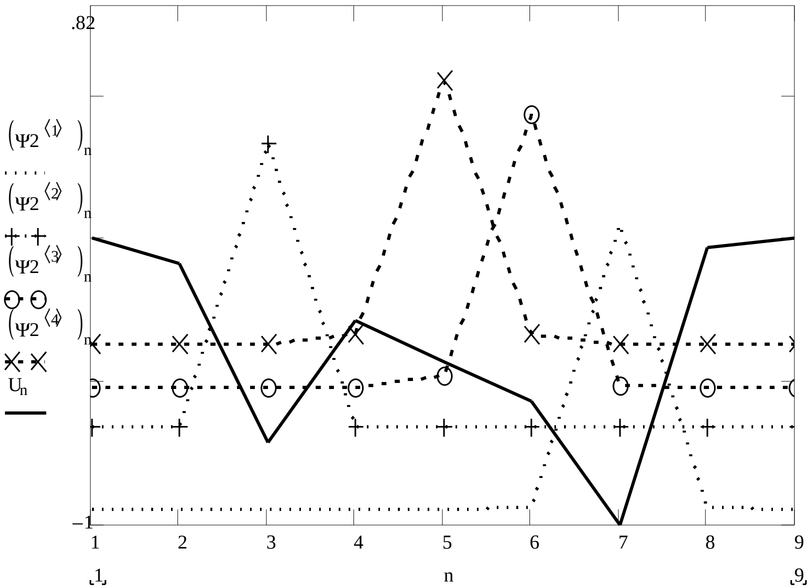

Рис. 2. Квадраты собственных функций для первых четырех собственных значений бесконечно глубокой потенциальной ямы с рельефом дна $U = U(x)$

В качестве второго модельного примера рассмотрим построение функции распределения объектов в одномерной потенциальной яме конечной глубины. Потенциал в этом случае определяется следующим образом

$$U = \begin{cases} 0, & \text{если } x \le 0 \qquad \text{(I)} \\ U(x), & \text{если } 0 \le x \le L \quad \text{(II)} \\ 0, & \text{если } x \ge L \qquad \text{(III)} \end{cases}$$

Такая яма может служить моделью потенциала с конечным радиусом действия, не оказывающим влияние на объекты за пределами этого радиуса.

Функция Ψ в этом случае должна удовлетворять однородному линейному дифференциальному уравнению с постоянными коэффициентами

$$\frac{d^2}{dx^2}\Psi + k^2\Psi = 0,$$

где $k = \sqrt{(E-U)}$.

На функцию Ψ, являющуюся решением этого уравнения, накладываются требования непрерывности и непрерывной дифференцируемости. Кроме этого, она должна быть однозначной и конечной во всем пространстве. Эти требования приводят к тому, что решения полученного уравнения существуют, вообще говоря, не при любых, а только при некоторых значениях параметра E. Такие значения называются собственными значениями задачи.

Вид решения полученного дифференциального уравнения существенно зависит от соотношения величин E и U. В случае, когда $E < U$, решение будет носить экспоненциальный характер, и из двух экспонент выбирается та, которая принимает нулевое значение на бесконечности. В случае, когда $E > U$, величина k будет мнимой, и следовательно, решение будет иметь колебательный характер, т.е. будет представлять некоторый волновой процесс.

Исходя из этих требований и накладываемых на функцию Ψ ограничений, решение в областях (I) и (III) ищется в виде $\Psi(x) = \Psi_0 e^{kx}$. В области (II) для нахождения решения дифференциального уравнения используется численный метод. Для этого уравнение заменяется системой дифференциальных уравнений

$$\begin{cases} p = \dfrac{d\Psi}{dx} \\ \dfrac{dp}{dx} = k^2\Psi, \end{cases}$$

которые затем в свою очередь заменяются системой разностных уравнений следующего вида

$$\begin{cases} p_{i+1} = p_i + k^2\Psi_i\Delta \\ \Psi_{i+1} = \Psi_i + p_i\Delta. \end{cases}$$

Решение в области (II) начиналось с граничного значения, полученного в области (I), и строилось посредством разностной схемы. Значение, полученное на границе области, сравнивалось с граничным значением в области (III). Величина параметра E подбиралась таким образом, чтобы абсолютная величина разности граничных значений функции Ψ и ее производной не превышала заданной точности ε.

На рис. 3 и 4 представлены результаты численных расчетов, выполненных в соответствии с предложенным подходом.

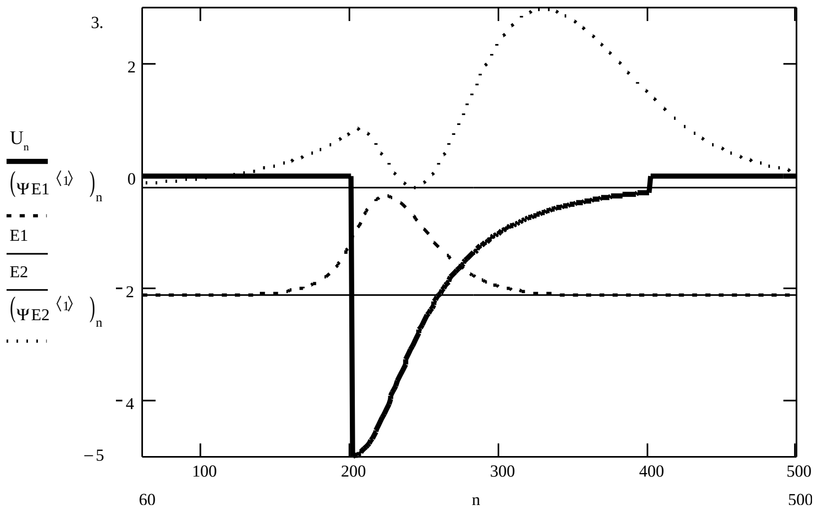

Рис. 3. Функции распределений плотностей первых двух собственных значений потенциала
$U(x) = -5/(1+x^2)$

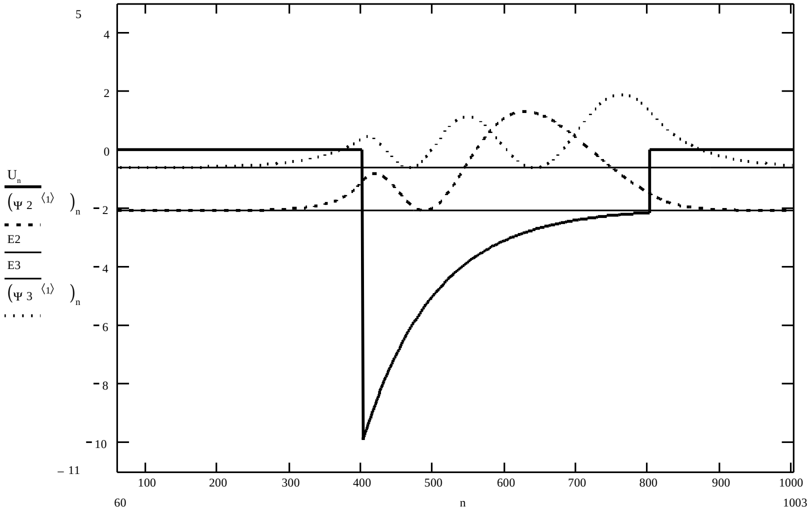

Рис. 4. Функции распределений плотностей второго и третьего уровней для потенциала вида $U(x) = -(8e^{-x}+2)$

Пример содержательной интерпретации классификационной задачи

В рамках сформулированного подхода рассмотрим задачу моделирования распределения населения России по величине среднедушевого дохода в сентябре-октябре 1996 г., предложенную в [8]. Согласно представленным результатам все население Российской Федерации было разделено на 5 отличающихся друг от друга по социальному положению и уровню доходов классов или, используя терминологию [8, с. 77], страт. Результаты, в части интересующей нас, воспроизведены в табл. 1. Графическое представление результатов статьи [8] воспроизведено на рис. 5.

В соответствии с предлагаемой нами моделью каждый индивид (физическое лицо – с точки зрения налоговых органов) будет представлен объектом в одномерном признаковом пространстве, в качестве координаты которого выступает доход этого индивидуума.

Будем предполагать, что каждый индивид обладает некоторой энергией, которая позволяет ему иметь ту или иную величину дохода. Возможность получать доход определяется, с одной стороны, его внутренней энергией (обозначим эту величину E), а с другой – сдерживается проводимой фискальной политикой, которая в нашей модели будет выполнять функцию потенциала $U(x)$.

Таблица 1

ОСНОВНЫЕ ХАРАКТЕРИСТИКИ СТРАТ РАСПРЕДЕЛЕНИЯ РОССИЙСКИХ СЕМЕЙ ПО СРЕДНЕДУШЕВОМУ МЕСЯЧНОМУ ДОХОДУ[1]

№ класса (страты)	Удельный вес доходной группы	Среднее значение среднедушевого мес. дохода (тыс. руб.)
1	0,3900	350
2	0,4000	700
3	0,1700	1640
4	0,0397	12000
5	0,0003	120000
Сумма	1,0000	–

Принимая во внимание, что обложение в рассматриваемый период проводилось по следующей шкале:

до 12000000 руб. 12%;
от 12000001 до 24000000 руб. 20%;
от 24000001 до 36000000 руб. 25%;
от 36000001 до 48000000 руб. 30%;
свыше 48000001руб. 35%,

будет очевидно, что $U=U(x)$. Графическое изображение этой схемы представлено на рис. 6 сплошной линией. По оси абсцисс отложена величина среднемесячного дохода в логарифмическом масштабе, а по оси ординат – величина налога в процентах.

[1] Фрагмент табл. 1 из [8, с. 80].

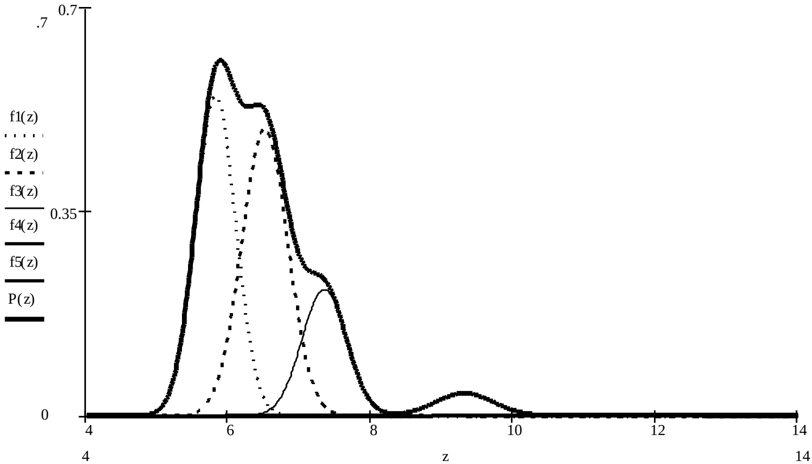

Рис. 5. Плотности распределения отдельных классов и смеси распределений по величине логарифма среднедушевого семейного дохода (воспроизведено по [8, с. 81]

Мы примем модель бесконечно глубокой ямы, поскольку, с одной стороны, величина дохода не может принимать отрицательных значений, а с другой, – величина $U(x)$ порядка $e^{20} \approx 4 \cdot 10^{8}$ минимальных месячных оплат труда (ММОТ). Такой доход, согласно официальным данным Госкомстата РФ (см. [9, с. 176]), сопоставим с общими денежными доходами населения страны в сентябре 1996 г., составившими 116 трл. руб., и вряд ли мог быть получен в качестве дохода физическим лицом. Поэтому для численного решения задачи выберем сетку с числом узлов, равным 20.

В качестве величины, характеризующей однородность признакового пространства, в рамках нашей модели естественным образом выступает величина ММОТ. Трудности, возникающие при рассмотрении этой модели, в значительной степени связаны с той степенью неопределенности данных, которая присуща экономическим процессам переходного периода. Так, в этот период несколько раз менялся размер минимальной заработанной платы.

В наших расчетах эта величина, точнее – логарифм этой величины, был принят равным 75900 руб.

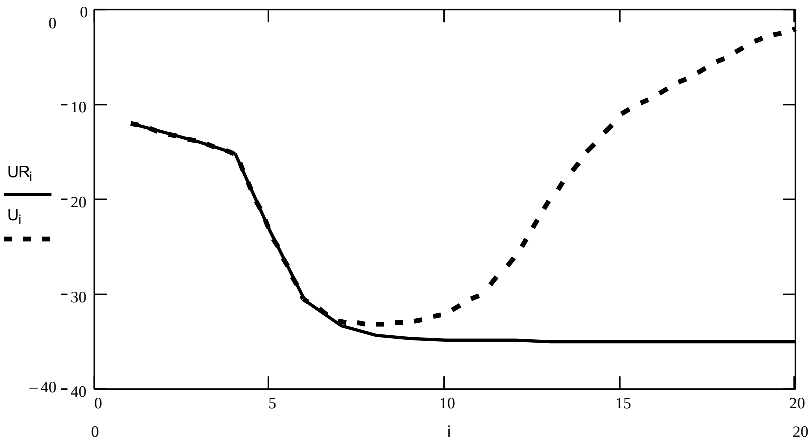

Рис. 6. Потенциал, соответствующий нормативным актам, – сплошная линия, потенциал, предполагаемый в модели, – пунктирная линия

Первоначальный расчет плотностей распределения классов, выполненный в расчете на то, что потенциал имеет форму, показанную на рис. 6, не позволяет получить распределение, даже близкое к распределению рис. 5. Поэтому предположим, что в реальности имела место ситуация, представленная на рис. 6 пунктирной линией, т.е. реальный потенциал, а с ним и вся схема налогообложения носила несколько другой характер.

Согласно предполагаемой схеме, реальное налогообложение соответствовало нормативным актам только в областях малых и средних доходов – примерно до 1100 тыс. руб., а при более высоких доходах происходило их сокрытие (укрывание). Причем, чем выше были доходы, тем меньшая их часть шла на выплату налогов. Так, например, при доходе 442413 тыс. руб. (e^{13}) величина уплаченного налога составила примерно 88428 тыс. руб., а при доходе 65659969 тыс.

руб. (e^{18}) – примерно 3282998 тыс. руб. Хотя в первом случае следовало уплатить 154844, а во втором – 22980989 тыс. руб.

Результаты численного решения стационарного уравнения Шредингера с выбранным потенциалом для однородных граничных условий представлены на рис. 7.

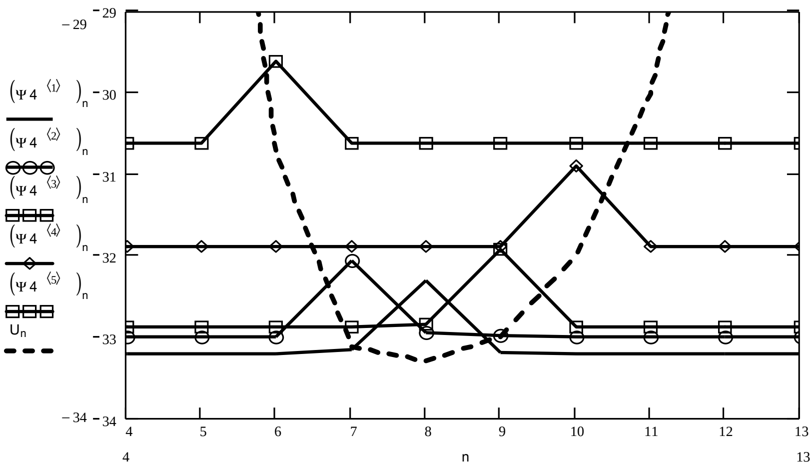

***Рис. 7. Квадраты первых пяти собственных функций уравнения. Пунктир – часть потенциала** $U(x)$*

Анализ квадратов собственных функций для первых пяти собственных значений показывает, что если считать, что классы упорядочены по величине энергии объектов, входящих в них, то возникают определенные трудности с подбором весов для функции, представляющей плотность распределения всей совокупности. Если же отказаться от такого предположения и считать, что классы ранжированы следующим образом: 3; 2; 4; 5; 1, то удается достигнуть довольно хорошего совпадения полученных результатов на качественном уровне. Это можно увидеть на рис. 8.

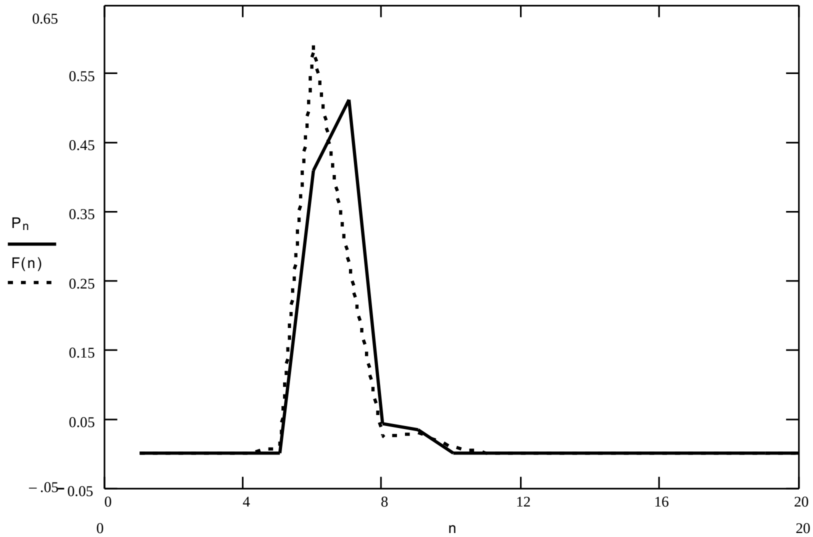

Рис. 8. Результаты нашего расчета – сплошная линия, результаты статьи [8] ***– пунктирная линия***

Анализируя полученные значения величины *E*, отметим, что согласно предложенной модели, наименьшей энергией (усилиями), затраченной на получение среднедушевого дохода, обладают представители класса № 3 – рядовые работники госаппарата и фирм и класса № 2 – работники бюджетной сферы и свободных профессий. По-видимому, в силу хороших «стартовых» позиций, социально-экономическое положение этих классов гарантирует им получение соответствующих доходов. А вот представителям классов № 4 – владельцам, основным акционерам и ключевым сотрудникам относительно преуспевающих отраслей экономики и № 5 – высшим слоям госбюрократической и мафиозной элит приходится расходовать более значительную величину энергии для получения своих доходов. И гораздо более напряженной (совсем) представляется по сравнению с предыдущими классами ситуация в классе № 1, который составляют безработные допенсионного возраста, неработающие пенсионеры, низкооплачиваемые наемные работники.

Используя полученные результаты и методику работы [8], можно оценить суммарную величину необъявленных (скрытых) доходов. Согласно нашей модели, эта величина составляет 82,5% против 54%, указанных в работе [8]. Возможно, расхождение вызвано неточностями в оценках и в уточнении нуждаются значения численностей доходных групп и величина минимальной месячной оплаты труда. При этом заметим, что в рамках предлагаемого подхода появляется возможность, как это было показано ранее, оценить величины скрытых доходов в отдельности по каждому из классов.

***

Задача классификации как задача изучения структуры многомерных данных в признаковом пространстве приводит к необходимости объяснения механизма формирования этой структуры, т.е. объяснения наличия классов в совокупности объектов, их взаимного расположения и численности. Было показано, что интерпретация этих структурных особенностей может быть получена за счет довольно общих предположений:

1. Все объекты рассматриваемой совокупности обладают некоторой способностью к изменению своего положения в системе.

2. Признаковое пространство не является однородным, в нем определен некоторый потенциал $U = U(x)$, который сдерживает произвольно большие отклонения объектов системы и формирует структуру системы объектов.

Выполнение этих предположений логически приводит к тому, что распределение элементов системы в признаковом пространстве может быть описано уравнением типа уравнения Шредингера. Интегрирование этого уравнения, при накладываемых на функцию Ψ естественных ограничениях конечности и непрерывной дифференцируемости, позволяет перевести связи из кинематических в статистические и объяснить наличие устойчивых состояний в системе объектов. Эти состояния, очевидно, могут быть соотнесены с эмпирическими классами, выделяемыми алгоритмами кластерного анализа.

Рассмотренный подход к классификационной задаче [10] открывает большие возможности для дальнейшего развития теории классификации. Так, сложные системы, согласно современным представлениям, характеризуются тем, что глобально детерминированы и в то же время локально случайны. Именно такой характер носит предлагаемая модель – наличие классов, их интенсивности и функции распределения, определяемые как решения уравнения Шредингера, являются детерминированными, а распределение объектов в классах носит стохастический характер. Наблюдаемая структура многомерных данных служит проявлением сочетания случайных и детерминированных законов существования рассматриваемой системы.

Общее решение введенного в рассмотрение уравнения может быть представлено в виде ряда

$$\Psi = \sum_n C_n \Psi_n,$$

и, принимая во внимание условие ортонормированности собственных функций задачи, можно видеть, что

$$(\Psi, \Psi') = \sum_n C_n^{\ 2}.$$

Это позволяет интерпретировать коэффициенты C_n^2 как интенсивности классов, т.е. величины, пропорциональные их численности. Требование конечности функции Ψ приводит к тому, что ряд, составленный из коэффициентов, должен быть сходящимся, и следовательно, $C_n^2 \to 0$. Предположим, что коэффициенты этого ряда в простейшем случае обратно пропорциональны порядковым номерам классов, т.е.

$$C_n^2 = \frac{B}{n^p},$$

где B, p – некоторые константы. Тогда для сходимости ряда, члены которого задаются таким соотношением, очевидно, должно выполняться требование $p > 1$, что автоматически приводит к тому, что ранговое распределение численности классов должно носить гиперболический характер. Выполнение это соотношения на эмпирических данных, как показано в [11], может служить необходимым условием системности изучаемой совокупности социально-экономических объектов.

ЛИТЕРАТУРА

1. Типология и классификация в социологических исследованиях. М.: Наука, 1982.

2. Интерпретация и анализ данных в социологических исследованиях. М.: Наука, 1987.

3. Математические методы анализа и интерпретация социологических данных. М.: Наука, 1989.

4. *Буховец А.Г.* Последовательное применение алгоритмов многомерной классификации // Многомерный анализ социологических данных (методические указания, алгоритмы и описания программ). М.: ИСИ АН СССР, 1981. С. 24–73.

5. *Татарова Г.Г.* Типологический анализ в социологии. М.: Наука, 1993.

6. *Фок А.В.* Начала квантовой механики. М.: Наука, 1976.

7. *Самарский А.А.* Теория разностных схем. М.: Наука, 1983.

8. *Айвазян С.А.* Модель формирования распределения населения России по величине среднедушевого дохода (экспертно-статистический подход) // Экономика и математические методы. 1997. Т. 33. Вып. 4. С. 74–86.

9. Социально-экономическое положение России. Январь-сентябрь 1996 г. / Госкомитет РФ по статистике. 1996. № 9.

10. *Буховец А.Г.* Квантово-механическая интерпретация задачи многомерной классификации // Обозрение прикладной и промышленной математики. М., 2001. Т. 8. Вып. 1. С. 120–121.

11. *Буховец А.Г., Соловьев А.С.* Критерий системности социально-экономических объектов // Математические методы в социологических исследованиях. М.: ИСИ АН СССР, 1984. С. 28–36.